Massachusetts Institute of Technology
6.435 Theory of Learning and System Identification

Prof. Dahleh, Prof. Mitter — Homework 4
Out 4/12 — Due F, 4/20

1 [Central Limit Theorem in the Asymptotics of the Estimator]

Consider the system identification setting that we presented in class. Our data is generated by a model of the form:

$$y_t = \sum_{i=1}^{n_a^\star} -a_i y_{t-i} + \sum_{j=0}^{n_b^\star} b_j x_{t-j} + w_t, \qquad w_t \text{ i.i.d. } \mathbf{E}[w_t] = 0,\ \mathsf{var}(w_t) = \lambda^2.$$

The memory or number of taps $n_a^\star$ and $n_b^\star$ are typically unknown, and we perform our optimization in a model class with some fixed n_a and n_b. In this context, minimizing the square regression cost on a data sample of size ℓ results in the following estimate:

$$\hat{\alpha}_\ell = \left(\frac{1}{\ell}\sum_{t=1}^{\ell} \phi_t \phi_t'\right)^{-1} \left(\frac{1}{\ell}\sum_{t=1}^{\ell} \phi_t y_t\right),$$

where $\phi_t = [-y_{t-1} \ \cdots \ -y_{t-n_a} \ x_t \ \cdots \ x_{t-n_b}]'$, and $\hat{\alpha}_\ell = [\hat{a}_1 \ \cdots \ \hat{a}_{n_a} \ \hat{b}_0 \ \cdots \ \hat{b}_{n_b}]'$. Assume the inverse exists, e.g. the inputs have sufficient persistence of excitation.

(a) Show that if $n_a^\star = n_a$ and $n_b^\star = n_b$ then:

$$\hat{\alpha}_\ell = \alpha^\circ + \frac{1}{\sqrt{\ell}} \underbrace{\left(\frac{1}{\ell}\sum_{t=1}^{\ell} \phi_t \phi_t'\right)^{-1}}_{\text{(I)}} \underbrace{\left(\frac{1}{\sqrt{\ell}}\sum_{t=1}^{\ell} \phi_t w_t\right)}_{\text{(II)}},$$

where α° represents the true parameters:

$$\alpha^\circ = [a_1 \ \cdots \ a_{n_a} \ b_0 \ \cdots \ b_{n_b}]'.$$

We are interested in the convergence behavior of the error $\hat{\alpha}_\ell - \alpha^\circ$. Let's consider the simple context of a one-step memory $n_a = 1$, $n_b = 0$. Under the usual ergodicity and quasi-stationarity assumptions, term (I) converges to some matrix P.

(b) Carry out the argument given in class and explicitly separate term (II) into components for which the central limit theorem applies, and show that it converges to a zero-mean Gaussian random variable with covariance matrix $\lambda^2 P$.

(c) Deduce that we have $\sqrt{\ell}(\hat{\alpha}_\ell - \alpha^\circ) \sim \mathcal{N}(0, \lambda^2 P^{-1})$. Provide a brief statement on what this says about the convergence of the estimator.

2 [Expectation Maximization]

Consider the EM algorithm for estimating the parameters of a hidden Markov model.

[E-step] $J(\tilde{\alpha}, \alpha) = \sum_{x^\ell} \left[\log \mathbf{P}_{\tilde{\alpha}}(Y^\ell = y^\ell, X^\ell = x^\ell)\right] \mathbf{P}_\alpha(Y^\ell = y^\ell, X^\ell = x^\ell)$.

[M-step] $\alpha^\star = \text{argmax}_{\tilde{\alpha}}\, J(\tilde{\alpha}, \alpha)$.

[Update] $\alpha \leftarrow \alpha^\star$.

Verify that the solution of the maximization step stated in class is correct:

$$\alpha^\star = (\Pi^\star, a^\star, b^\star),$$

Such that :

$$\begin{aligned}
\Pi^\star(i) &= P_\alpha(Y^\ell = y^\ell, X_0 = i)/P_\alpha(Y^\ell = y^\ell), \\
a^\star_{ij} &= \sum_{t=1}^{\ell} P_\alpha(Y^\ell = y^\ell, X_{t-1} = i, X_t = j)/\sum_{t=1}^{\ell} P_\alpha(Y^\ell = y^\ell, X_{t-1} = i), \\
b^\star_i(v) &= \sum_{\{t\,:\,y_t = v\}} P_\alpha(Y^\ell = y^\ell, X_t = i)/\sum_{t=1}^{\ell} P_\alpha(Y^\ell = y^\ell, X_t = i).
\end{aligned}$$

3 [Recursive Filtering]

Given a hidden Markov model, define the sequence of vectors q_t:

$$q_t(i) = \mathbf{P}(X_t = i | Y^t = y^t), \qquad i = 1, \cdots, n,$$

which is the posterior distribution of state X_t given all the observations up to that time. Show that $q_{t+1} = \Phi(q_t, y_{t+1})$, for some Φ, i.e. that q_t satisfies a recursion.

4 [Decoding via the Viterbi Algorithm]

Given a hidden Markov model, with known parameters, and an observation sequence y^ℓ, the decoding problem consists of computing the complete sequence of states with maximal posterior probability:

$$\hat{x}^\ell = \underset{x^\ell}{\text{argmax}}\ \mathbf{P}(X^\ell = x^\ell | Y^\ell = y^\ell).$$

(a) Define the sequence of vectors r_t as follows:

$$r_t(i) = \max_{x^{t-1}} \mathbf{P}(X_t = i, X^{t-1} = x^{t-1} | Y^t = y^t).$$

Show that r_t satisfies a forward recursion such that:

$$r_t(i)\ \alpha\ b_i(y_t) \max_j a_{ji} r_{t-1}(j),$$

where the proportionality constant depends only on the observations and t, but is independent of the optimization.

(b) Suppose r_t, $t = 1, \cdots, \ell$ are all computed, up to the proportionality factor. Show that, if we also keep track of $s_t(i)$:

$$s_t(i) = \underset{j}{\text{argmax}}\, a_{ji} r_{t-1}(j),$$

then the optimal sequence $\hat{x}^\ell$ can be recovered by a backward recursive selection.